OPTIMIZING HUMANITARIAN LOGISTICS WITH DEEP REINFORCEMENT LEARNING AND DIGITAL TWINS

Bulent Soykan[a] and Ghaith Rabadi[a]

[a] School of Modeling, Simulation, and Training, University of Central Florida, Orlando, FL, USA
{bulent.soykan, ghaith.rabadi}@ucf.edu

ABSTRACT

This paper explores the optimization of humanitarian logistics, with a focus on the Multi-Commodity Network Flow (MCNF) problem in this setting. Our primary goal is to enhance the efficiency of aid distribution, minimizing long-term transportation costs while managing the complexities of demand and supply in disaster/emergency situations. To accomplish this, we suggest a deep reinforcement learning (DRL) method that incorporates graph neural networks to approximate the value function. Also, our approach involves the use of digital twins (DTs) for accurate modeling and simulation, reflecting the dynamic and stochastic nature of humanitarian logistics. Our computational experiments include a comparative analysis against traditional deterministic and heuristic methods. We examine the performance of our DRL approach across simulated MCNF problem instances. The results indicate DRL agent's capability in optimizing logistics tasks, and the incorporation of DTs and DRL demonstrates effectiveness and adaptability in managing the humanitarian logistics.

Keywords: deep reinforcement learning, digital twins, graph neural networks, humanitarian logistics optimization, stochastic network flows.

1 INTRODUCTION

Humanitarian logistics have a vital role in effective disaster/emergency management, where the timely and efficient delivery of relief items is extremely important [1, 2]. However, this field faces a number of challenges that make it different from commercial logistics [3]. Humanitarian crises, such as natural disasters and conflicts, create rapidly changing scenarios where the demand for relief goods is uncertain and varies greatly over time [4]. This requires a logistics strategy that is adaptable and responsive to these changes. Also, the complexity of coordinating among diverse stakeholders – including international organizations, government agencies, NGOs, and local communities – each with its unique priorities and procedures, adds another layer of complexity to the execution of humanitarian logistics.

The challenges in humanitarian logistics are further complicated by resource limitations and the diverse nature of relief items [5]. Unlike commercial operations, humanitarian efforts often face strict budget constraints and a limited availability of transportation resources. In order to ensure efficient relief distribution in this situation, it is imperative that resources be used as efficiently as possible. Moreover, humanitarian aid comprises a wide range of commodities, such as food, water, medical supplies, and shelter materials, each requiring specific storage and transportation conditions. The need to manage these multi-commodity flows efficiently, coupled with challenges posed by damaged infrastructure and accessibility issues in affected areas, necessitates innovative logistics solutions to make sure aid reaches the most vulnerable. These distinct challenges require a flexible, robust, and adaptive approach where traditional deterministic methods are often inadequate in addressing the dynamic and complex nature of humanitarian logistics.

Proc. of the 2024 Annual Simulation Conference (ANNSIM'24), May 20-23, 2024, American University, DC, USA
C. Ruiz-Martin, B. Oakes and R. Cárdenas, eds.

In these crisis situations, the main goal for the optimization of transportation and distribution networks is to deliver relief goods efficiently and timely from donors/storage centers to those affected by disasters or emergency situations. This task is specifically complicated by variable demand and the need to consolidate shipments. These challenges are amplified by the uncertainty inherent in emergencies, making the prediction of needs and the coordination of logistics an intricate real-life problem. Moreover, the optimization of transportation in humanitarian logistics should also focus on cost-effectiveness. With finite resources and often stringent budgetary restrictions, humanitarian organizations must ensure that every dollar spent has the maximum possible impact. This efficiency is not only a financial consideration but also a moral imperative. By optimizing routes, consolidating loads, and managing the flow of multiple types of commodities, organizations can serve more affected individuals. Moreover, efficient logistics can prevent the wastage of vital resources, allowing for sustained support over the longer term, which is often crucial in recovery efforts following initial emergency response.

Furthermore, the optimization of transportation and distribution has broader implications beyond immediate cost and speed benefits. It enhances the reliability and trust in humanitarian operations, which is crucial for ongoing support from donors. An optimized logistics system should adapt to changing conditions, maintaining the flow of aid even when the situation on the ground changes. The integration of transformative technologies including intelligent digital twins (DTs) and machine learning holds significant potential for these optimization efforts. Such innovative approaches can lead to more adaptive, more responsive logistics models that can handle the complexities of real-world humanitarian crises.

The Multi-Commodity Network Flow (MCNF) problem is one of the main fundamental problems in humanitarian logistics, representing the complexity of managing multiple streams of aid goods through a network of logistical pathways [6]. This problem encapsulates the challenge of simultaneously transporting different types of commodities from a variety of donation/aid storage centers to numerous demand locations. The MCNF problem is inherently multi-stage and stochastic, reflecting the unpredictable and evolving nature of crises. Addressing this real-world problem requires a sophisticated understanding of network dynamics and the ability to make decisions that account for the interplay between different commodities and the constraints of the network. Traditional approaches often fall short in such a complex environment, unable to capture the stochastic nature of demand and the intricacies of multi-commodity flows.

In this paper, we propose an approach for optimizing the MCNF problem within the context of humanitarian logistics with deep reinforcement learning (DRL) and DTs in order to address the outlined challenges. In particular, our approach consist of development of a DRL agent with graph neural networks (GNNs) as the value function approximator, and integration of a DT for a realistic simulation of humanitarian logistics networks. Our main research question is: How does a DRL agent with a GNN component, enhanced by a DT for simulation and predictive analytics, perform compare to traditional greedy heuristic and deterministic optimization policies in humanitarian logistics, focusing on metrics including cost reduction per consignment, on-time delivery rates, and reduced inter-facility transfers for aid demands?

This paper makes several scientific contributions to the field of computational logistics. First, we introduce a transformative approach that combines DRL with DTs for improving adaptability in logistics networks. This contribution is important for its potential to transform how logistical challenges are tackled in highly volatile humanitarian contexts. Second, we develop a comparative analysis framework that evaluates the DRL agent against traditional methods. This aspect of our research addresses the need for empirical evidence on the effectiveness of new optimization techniques in real-life settings.

The remainder of this paper is organized as follows: Section 2 introduces the MCNF problem in humanitarian logistics context, reviews traditional optimization methods and their limitations, and then explores how DRL, GNNs, and DTs address these challenges, providing context for the research's challenges and innovations. Section 3 is dedicated to the methodological details of our study, including the development of the DRL agent and the integration of DTs for real-time simulation purposes. Section 4 presents the results obtained from comparing our approach to traditional optimization methods, discussing

the performance and adaptability of our approach in simulated scenarios. Finally, Section 5 provides a recap of the findings and proposes potential avenues for future research in this evolving field.

2 BACKGROUND AND RELATED WORK

2.1 Multi-Commodity Network Flow Problem in Humanitarian Logistics

The MCNF problem involves routing multiple types of commodities through a single logistics network, each originating from different supply points and destined for various demand nodes [7]. This complexity is compounded by the diverse requirements of different commodities, such as specific handling, storage needs, and varying priorities, which must be managed within the constraints of the same logistical network. One of the seminal works in this field, presented by [8], highlights the inherent difficulties in finding efficient solutions to multi-commodity flows, especially given the often polynomial-time algorithms required for static multi-commodity flow calculations. These complexities are further amplified in dynamic environments typical of humanitarian logistics, where time factors play a crucial role. The introduction of time elements into network flows, as discussed by [9], transforms the problem into a more complex dynamic flow problem. This dynamic aspect is crucial in humanitarian logistics, where the timely delivery of aid can be as critical as the aid itself.

The MCNF problem in humanitarian logistics setting also encompasses network constraints including limitations on transportation capacities, route availabilities, and storage capabilities at different network nodes. Moreover, each commodity type might impose its own specific constraints, like perishability or special handling requirements. In their study addressing disaster relief management, [10] present a formulation and two heuristic solutions for a large-scale multi-commodity, multi-modal network flow problem with time windows, utilizing a time-space network structure and evaluating the performance over various problem sizes using synthetic data sets.

Additionally, addressing the MCNF problem involves optimizing costs in resource-limited humanitarian operations, balancing transportation costs, timely delivery, and reliability against network constraints. In their investigation of the location and allocation problem within the MCNF problem, [11] present a mathematical optimization model to incorporate additional cost components, specifically production and holding costs at both the supply center and distribution center levels. They utilize a Particle Swarm Optimization algorithm enhanced with multiple social learning terms and they conduct a comparative analysis based on benchmark data sets. The study measures the impact of these cost factors on total cost, solution quality, coefficient of variation, and computational time, providing insights into the significance of including these elements in mathematical formulations.

2.2 Traditional Optimization Methods

Traditional optimization methods typically fall into two categories: deterministic and heuristic approaches [12]. Deterministic methods rely on predefined parameters and models to forecast demand and optimize logistic operations. They are characterized by their reliance on historical data and the assumption that future events will mirror historical trends. These methods are adept at solving linear and well-defined problems where the variables and outcomes are predictable. However, they often fall short in the volatile environments characteristic of humanitarian crises, where demands are uncertain and conditions change rapidly. Heuristic approaches, on the other hand, offer more flexible solutions. These methods employ rules of thumb and simplified decision-making processes to find good-enough solutions under uncertainty. Heuristics are particularly useful when the problem is too complex for an exact mathematical solution or when quick decisions are necessary. Examples include the use of greedy algorithms, which make locally optimal choices at each step, and simulation-based methods, which can model different scenarios to identify satisfactory solutions. Nevertheless, while heuristics are more adaptable than deterministic methods, they may not always yield the most optimized solution and can lead to suboptimal performance when faced with the multi-faceted challenges of humanitarian logistics. The unpredictable, dynamic nature of humanitarian

emergencies and the complexity of managing multiple commodities call for a more robust and adaptive optimization strategy, highlighting the limitations of traditional methods that cannot accommodate real-time data or rapidly adapt to environmental changes [13].

2.3 Deep Reinforcement Learning, Graph Neural Networks And Digital Twins

Reinforcement Learning (RL) algorithms learn optimal actions based on feedback from their environment. This learning process is particularly effective in dynamic and uncertain scenarios, like those encountered in humanitarian logistics. RL's strength lies in its ability to adaptively make decisions in response to changing conditions. RL approaches are typically categorized into value-based and policy-based methods. Value-based methods, including Q-learning, focus on determining the value of being in a state and taking an action, whereas policy-based methods directly learn the policy without valuing states. Q-learning, a value-based method, is commonly used because it effectively balances the trade-off between exploration and exploitation, which is crucial in uncertain environments. In Q-learning, the agent explores the environment and learns a policy by updating a table of Q-values, which estimate the total future rewards for each action in each state. By means of iterative interactions with the environment, the agent optimizes these Q-values, ultimately reaching an optimal policy that maximizes the expected cumulative reward. The Q-value updates are based on the Bellman equation, which recursively decomposes the future reward. The agent selects actions by consulting the Q-table and typically employs a strategy that balances exploration of new actions with exploitation of known rewarding actions. This balance is critical in dynamic environments to prevent the agent from getting stuck in suboptimal policies [14].

However, as the complexity of the problem increases, Q-learning becomes impractical due to the exponential growth of the Q-table, which can neither be stored nor efficiently updated. This is where Deep Q-Network (DQN), which combines Q-Learning with deep neural networks, comes into play. DQN leverages neural networks to approximate the Q-table, effectively dealing with large or continuous state spaces that are characteristic of complex problems such as the MCNF problem. The neural network, trained to predict Q-values, allows the agent to generalize from observed states to unseen states, making it robust to the variance in real-world scenarios. Figure 1 demonstrates the capabilities of DQN in assessing various possible actions for any given state. In the upper section of the figure, the Q-learning approach is depicted, characterized by its generation of a single Q-value for each state-action pair. Conversely, the lower section of the figure showcases DQN's approach, which contrasts by producing an array of Q-values for a range of potential actions, all within the same state. This distinction highlights DQN's ability to navigate and evaluate complex decision-making scenarios.

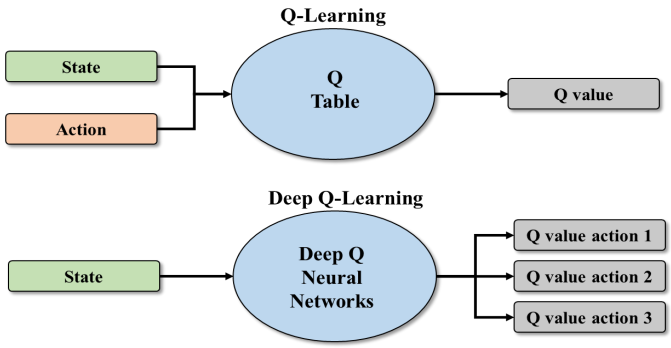

Figure 1: Q-learning and Deep Q-learning.

GNNs are adept at processing graph-structured data, making them particularly suitable for network flow problems. By representing transportation networks as graphs, GNNs have the potential to effectively manage the interdependencies and constraints inherent in the MCNF problem. In the context of DRL, integrating GNNs as the value function approximator can significantly enhance the capability of the learning agents [15]. While DRL can optimize decision-making processes by learning from interactions

with the environment, GNNs can provide a deeper understanding of the structured network data, which is often crucial in logistics and transportation settings. By leveraging the spatial relational information encoded in graphs, a GNN can help the DRL agent to better understand and navigate complex network structures, leading to more efficient and effective policy decisions [16].

The next critical component in our approach is the DT which is incorporated for real-time simulation and modeling. DTs are virtual replicas of physical systems that simulate real-world conditions [17], and integrate real-time data, ensuring that physical systems are continuously improved and remain effective in dynamic conditions [18]. In the context of supply chains, DTs enable the modeling of transportation networks and the dynamic scenarios [19]. DTs provide a comprehensive training platform for DRL algorithms, enabling them to learn and adapt to a variety of situations with high fidelity. DTs also offer a safe and cost-effective means for testing DRL strategies, eliminating the risks and expenses associated with physical world implementations. DTs enable quick iterative refinements of strategies through rapid scenario simulations, accelerating the development of more effective solutions.

3 MATERIALS AND METHODS

3.1 Problem Statement and Mathematical Formulations

A critical component of the humanitarian logistics network is the transportation of relief goods from donation and aid storage facilities, like warehouses, to affected areas (demand nodes) efficiently and timely. Considering that aid packages often include various items and these items might be stored across different locations due to factors like proximity to aid supply points or strategic pre-placement, inventory management becomes increasingly challenging as the number of commodities and recipient locations increases. Selecting the most suitable donation/storage center as the dispatch point for each consignment, when multiple centers can meet the needs of an affected area, can reduce transportation costs and minimize costly inventory transfers between donation and aid storage centers.

To address these logistics challenges, we formulate the MCNF problem in humanitarian logistics context as a mixed integer linear program (MILP), adapting the traditional MCNF formulation to include additional constraints to model the interaction between arcs leading to the same affected area. The MILP model is designed to capture the complex decision-making process involved in routing multiple commodities from different storage locations to affected areas. Let $G(N, A)$ represent a directed graph, where N is the set of nodes, A is the set of arcs, and $t=1,\ldots,T$ represent the time stages in the planning horizon. Each node in N can be a source, transit, or destination node for commodities $k \in K$, where K is the set of all commodities. The goal is to route a set of K commodities through the network G at the least possible total cost while satisfying flow-conservation constraints at the nodes. In the humanitarian logistics context, the physical network is expanded and duplicated for each type of aid commodity. First, if an arc from a storage center to an affected area has positive flow for any commodity, the solution must also exhibit positive flow for all commodities from that facility to the affected area and zero flow if there is no flow for any commodity. Second, if any arc from a given facility have positive flow, all corresponding arcs from other facilities to that affected area must exhibit zero flow. To incorporate consignment consolidation constraints, we introduce binary variables y that take the value 1 if a consignment for commodity k is consolidated at node i and shipped to node j, and 0 otherwise.

Parameters:

c_{ijt}^{k}: Transportation cost for shipping commodity k along the arc (i,j) in time period t.

d_{it}^{k}: Demand for commodity k at node i in time period t.

Q_{it}^{k}: Available quantity of commodity k at node i in time period t.

U_{ijt}^{k}: Maximum transportation capacity of commodity k along the arc (i,j) in time period t.

Decision variables:

x_{ijt}^{k}: Amount of commodity k to be transported along the arc (i,j) in time period t.

y_{ijt}^{k}: Binary variable for ensuring that is 1 if a consignment for commodity *k* is consolidated at node *i,* and shipped to node *j* in time period *t,* and 0 otherwise.

$$Min\ Z\ =\ \mathbb{E}\left(\sum_{t=1}^{T}\sum_{k\in K}\sum_{i,j\in A} c_{ijt}^{k}x_{ijt}^{k}\right), \tag{1}$$

subject to

$$\mathrm{x_{ijt}^{k}} \le \mathrm{U_{ijt}^{k}},\quad \forall \mathrm{k} \in \mathrm{K},\ \forall \mathrm{t} = 1,\dots,\mathrm{T},\ \forall(\mathrm{i},\mathrm{j}) \in \mathrm{A}, \tag{2}$$

$$y_{ijt}^{k}\cdot Q_{it}^{k} \le x_{ijt}^{k} \le U_{ijt}^{k}\cdot y_{ijt}^{k},\quad \forall k \in K,\ \forall t = 1,\dots,T,\ \forall(i,j) \in A, \tag{3}$$

$$\sum_{k\in K}\sum_{i\in N} y_{ijt}^{k} \le 1,\quad \forall j \in N,\ \forall t = 1,\dots,T, \tag{4}$$

$$\sum_{\mathrm{j:(i,j)\in A}} \mathrm{x_{ijt}^{k}} - \sum_{\mathrm{j:(j,i)\in A}} \mathrm{x_{jit}^{k}} = \mathrm{d_{it}^{k}},\quad \forall \mathrm{i} \in \mathrm{N},\ \forall \mathrm{k} \in \mathrm{K},\ \forall \mathrm{t} = 1,\dots,T, \tag{5}$$

$$x_{ijt}^{k} \ge 0,\ y_{ijt}^{k} \in \{0,1\},\quad \forall k \in K,\ \forall t = 1,\dots,T,\ \forall(i,j) \in A. \tag{6}$$

The objective is to minimize the expected total cost of transportation over the entire planning horizon, considering the uncertainty in demand, given by (1). (2) ensures that the transportation capacity constraints are not exceeded at any time. This constraint mandates that the quantity of each commodity transported along any arc in the network, at any given time, must not surpass the maximum capacity of that arc for the commodity. (3) addresses the consignment consolidation constraints. This is a critical aspect in humanitarian logistics, where aid consignments need to be dispatched from the same location to ensure efficiency and minimize handling. The equation enforces that a consignment for a commodity is consolidated at a node and shipped to a destination with positive flow. (4) outlines the exclusive arc usage constraints. In a practical sense, this constraint avoids scenarios where multiple facilities send partial shipments of the same commodity to the same affected area. Such a situation could lead to inefficiencies and increased transportation costs. Therefore, if one center starts shipping a specific commodity to an affected area, other centers are restricted from shipping the same commodity to the same area within the same time period. (5) ensures that for any given node (except where commodities are strictly entering or leaving, such as in source or destination nodes), the total incoming flow of a commodity is equal to the total outgoing flow of that commodity. (6) imposes non-negativity and integrality constraints on the decision variables. These constraints ensure that the quantities of commodities transported are physically feasible (non-negative) and adhere to the binary nature of the consolidation decisions (either a consignment is made or not).

Considering the fact that number of variables (locations, commodities, time periods) is quite large in humanitarian logistics, the computational resources required to solve this stochastic MILP model is prohibitively large. Also, this model is static, meaning it doesn't inherently adapt to changes over time. In a rapidly changing situation like a humanitarian crisis, the ability to adapt to new information in real-time is important. While this model incorporate uncertainty to some degree, it does so in a limited or linearized fashion. However, real-world scenarios involve complex and non-linear uncertainties that are hard to model accurately in this framework. Also, this model focuses on optimizing immediate outcomes and may not adequately capture the long-term impacts of decisions, which are crucial in humanitarian logistics for sustainable and efficient resource allocation.

While the stochastic MILP model offers a structured approach to optimization, their application is limited in dynamic, complex, and uncertain environments. Markov Decision Processes (MDPs), on the other hand, provide a more flexible and adaptive framework for sequential decision making that is better suited for the real-time, long-term, and stochastic nature of humanitarian logistics challenges. Therefore, we model the problem as an MDP and apply RL methods to find an optimal policy for aid distribution. An MDP is defined by its states, actions, transition probabilities, and rewards. We mapped each component of the MDP to the humanitarian logistics context as follows:

- States (S): A state in the MDP represents the current configuration of the humanitarian logistics network at a given time. It includes the inventory levels at each center, the status of current aid requests and the condition of the transportation network.
- Actions (A): Actions are decisions made by the agent including assigning a particular consignment to a center, choosing a transportation route and reallocating resources within the network.
- Transition Probabilities (P): These probabilities define the likelihood of moving from one state to another after taking a particular action. In our setting, this involves the probability of successfully delivering aid given the chosen route and the likelihood of changes in demand for aid.
- Rewards (R): The reward function provides feedback to the agent about the value of the actions taken. In our setting, rewards are defined based on the efficiency and effectiveness of aid distribution as the costs saved and the speed of delivery.
- Discount Factor (γ): In MDP, future rewards are discounted to reflect the preference for immediate rewards over future ones. In our setting, this represent the urgency of delivering aid, where immediate distribution is more valuable than future distribution.

3.2 Simulation Environment

We developed a simulation environment to provide a virtual testing ground for the DRL agent. This environment is a crucial component of our approach, as it allows for the training and evaluation of the agent. The simulation replicates the complexities and dynamics of humanitarian logistics operations, offering a platform where the DRL agent can interact, learn, and adapt its strategies. In the training phase, the DRL agent interacts with the simulation environment and learns to identify patterns, make predictions, and choose actions that optimize the logistics objectives. The simulation provides a dataset of experiences, from which the agent can learn and improve its decision-making algorithms. To ensure the simulation reflects the real-world, we designed it to emulate the logistics network which encompass all donation/aid storage centers, aid distribution centers, and affected areas (demand locations). Each node and connection within this network has a set of possible actions that can be executed. These actions include sending aid from one node to another, replenishing supplies at a center.

In the simulation environment, we designed several functions to model different aspects of humanitarian logistics. The aid request generation function is designed to emulate the volatile demand for aid that arises in affected regions. This function considers various factors such as the severity of the crisis, population density, and the extent of the disaster's impact to generate realistic aid requests from different areas within the network. The inventory generation function is designed for modeling the supply of aid commodities available at different centers within the network. It dynamically simulates inventory levels based on several inputs: the existing inventory level, the list of aid requests. This function ensures the simulation environment reflects the fluctuating availability of aid supplies and the logistical challenge of maintaining adequate inventory levels. The optimization function is utilized to achieve the most cost-effective distribution of aid. It takes the graph representation of the logistics network and calculates the optimal routes for aid commodities to travel through the network. This function aims to minimize the overall costs of distribution while meeting the demand at various affected areas, considering constraints like transportation costs, route capacities, and the urgency of aid delivery. A dynamic element of the simulation is captured by expanding the logistics network over multiple time periods, creating a detailed model that accounts for the changing conditions of the network over time. This functionality allows for the modeling of temporal dynamics, such as the depletion and replenishment of aid supplies, evolving aid requests, and the operational impact of logistical decisions made in previous time steps.

Algorithm 1 describes the sequential process of the simulation environment. The simulation initiates by setting an initial state, where the agent selects an action to take within the network. Following this, the simulation enters a loop that only terminates upon meeting a specific end condition. Within this loop, the simulation updates the set of open aid requests, adjusts the inventory levels based on these requests and other changes, and transitions to a new state reflecting these updates. The agent's action is then evaluated to determine its success, offering a reward used to update the agent's decision-making policy. This new

state becomes the current state for the next iteration, where the agent selects another action, and the cycle repeats, processing each new action to continually refine the aid distribution strategy within the simulated network. This continuous loop allows the agent to learn and adapt to the evolving conditions of the network, striving to optimize the distribution of aid effectively and efficiently. The process is designed to mimic the complexities of real-world humanitarian logistics, where each decision can have significant and far-reaching consequences on the overall effectiveness of aid delivery.

Algorithm 1

```
current_state = get_initial_state() // Initialize the simulation environment with the initial state
action = agent.select_action(current_state) // Agent selects an action based on the current state
process_action(action) // Process the selected action in the simulation environment
while not is_simulation_end(): // Enter the main loop of the simulation, which continues until a termination condition is met
    update_open_requests() // Update the set of open aid requests in the network
    update_inventory() // Update the inventory levels at various storage facilities in the network
    new_state = get_current_state() // Update the current state to reflect the latest changes
    reward = evaluate_action(action, new_state) // Evaluate the effectiveness of the previous action and calculate the reward
    agent.update_policy(current_state, action, reward, new_state) // Update the agent's policy based on the action's outcome
    current_state = new_state // Set the current state to the new state for the next iteration
    action = agent.select_action(new_state) // Agent selects a new action based on the updated state
    process_action(action) // Process the new action in the simulation environment
```

3.3 Deep Reinforcement Learning Agent

DRL agent is the main entity programmed to make decisions regarding the dispatch of aid. The agent interacts with the simulation environment, processing current information about demand, supply levels, and network conditions to determine the best courses of action. Its decision-making process is informed by a policy that it continually refines as it learns from the outcomes of its actions. We used a replay buffer to store the experiences which is central to training the agent. The replay buffer, comprising states, actions, rewards, and subsequent states, with a predefined capacity, serves as a store of prior experiences, enabling the agent to gain knowledge from a variety of historical data. This mechanism is particularly advantageous as it enables the agent to break the temporal correlations in sequential observations, thus facilitating more stable and efficient learning. Sampling from this buffer is a critical operation. In typical scenarios, the buffer randomly selects a subset of experiences, ensuring that each training batch is a representative mix of the stored experiences. This random sampling aids in breaking correlations in the sequence of experiences, contributing to the robustness of the learning process.

The DRL agent employs GNNs to approximate the value function by creating embeddings for the entire logistics network using a Graph Convolutional Network (GCN). The GCN operates by processing the graph's data, generating embeddings for individual nodes within the network. These embeddings are then pooled to form a unified network representation. This pooled data is input into a fully connected layer, which outputs Q-values for each potential action, indicating their expected utility. The agent utilizes these Q-values to determine optimal actions for aid distribution, ensuring decisions stem from a comprehension of the network's state. This methodology guarantees that the agent's decision-making is rooted in a thorough understanding of the current state of the network, considering both individual node characteristics and their collective dynamics.

3.4 Integration of the Digital Twin

A custom DT is developed to mirror the complex network of logistics operations, providing a real-time, interactive, and adaptable framework for testing and refining the DRL strategies. This DT integrates several

data sources and utilizes modeling techniques to create a realistic and dynamic representation of the humanitarian logistics network. This digital representation allows the DRL agent to train in an environment that closely mimics real-world conditions, enabling it to develop strategies that are both effective and practical in real-life scenarios. The DT's adaptability ensures that the DRL agent is trained on the most current data, reflecting ongoing changes in the logistics network, such as new aid requests, shifts in resource availability, or changes in transportation infrastructure. This continuous updating process helps the DRL agent to stay relevant and effective, even as the real-world situation evolves.

Figure 2 illustrates the integration of the DT for adaptive decision-making. Physical Network, which represents the tangible elements of the real-world operations. Complementing this is the Historical Database, a repository that archives data from past operations, providing a rich source of information for pattern analysis and strategic planning. DT consists of several interconnected components: Data Processing component pre-processes both real-time and historical data. It ensures the information is clean, normalized, and formatted into an intricate data structure that mirrors the complexity of the real-world processes, complete with their interconnections and dependencies. Simulation Engine component simulates the real system and serves as a testing ground for various scenarios and predictions. This engine enables the system to forecast outcomes and explore the consequences of decisions before they are enacted in the physical world. The DRL agent receives input (system state and reward) from the Simulation Engine and is responsible for making decisions aimed at optimizing system's performance. As the DRL agent interacts with the Simulation Engine, it learns from both successful outcomes and errors through trial and error. The DRL agent continually receives feedback from the system's performance and uses this information to further improve its decision-making capabilities. This feedback loop allows the DRL agent to adapt to changes in the real system dynamically.

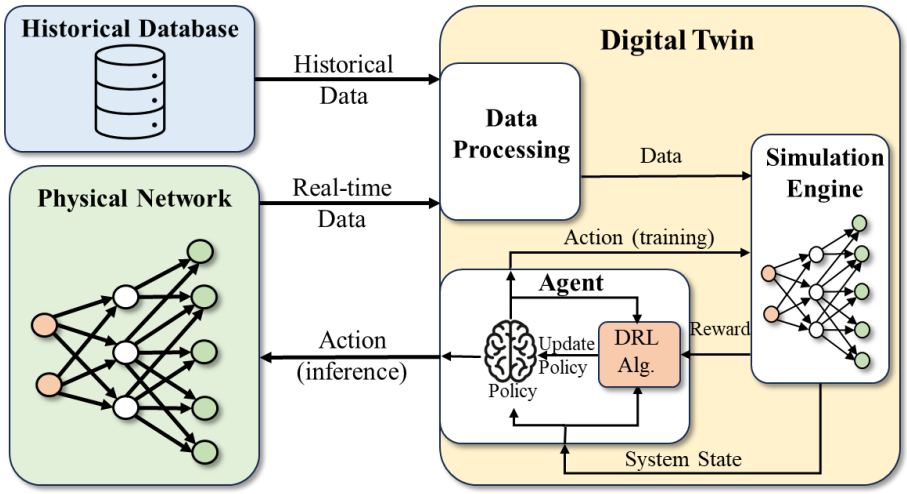

Figure 2: Integration of the Digital Twin.

3.5 Comparative Analysis Methodology

A comparative study is designed to evaluate the performance of our approach against traditional deterministic and heuristic optimization methods, and it is structured around several key performance metrics that are critical in the context of humanitarian logistics. These metrics are the reduction of average cost per consignment, the increase in the percentage of on-time deliveries, and the decrease in the frequency of inter-facility transfers required to meet aid demands.

We first establish baselines using greedy heuristic-based policies and deterministic optimization-based policies in order to achieve an objective comparison. Next, we utilize our simulation environment to generate a variety of scenarios that accurately reflect the dynamic and unpredictable nature of humanitarian crises. These scenarios are instrumental in testing both the traditional methods and our DQN agent. In each

scenario, we measure the performance of both traditional methods and the DQN agent. This measurement focuses on how each approach manages the complexities and challenges inherent in the MCNF problem. A critical aspect of our methodology is the statistical analysis conducted to evaluate the significance of the performance differences between the DQN agent and traditional methods. This step is crucial for validating the effectiveness of our proposed approach. For data collection and analysis, we rely on the simulation environment detailed in Section 3.2. This data is processed with analytical tools to ensure the accuracy and reliability of our performance measurements.

4 COMPUTATIONAL EXPERIMENTS, RESULTS AND DISCUSSION

4.1 Experimental Setup

Our computational experiments were structured to include various simulated disaster scenarios, reflecting the unpredictability and challenges of humanitarian logistics. These scenarios varied in terms of aid request uncertainty, supply constraints, and transportation difficulties. For simplicity, all transportation costs were fixed, and inventory costs were kept lower to minimize unnecessary inventory movement and associated costs. Inefficient allocation of inventory to wrong warehouses could lead to increased transportation costs and inefficient flow between routes. We implemented three different agents for the experiments: a greedy heuristic agent, a deterministic optimizer agent, and a DRL agent with GNN. The heuristic agent programmed to follow a strategy of selecting the most suitable center for dispatch based the proximity to the demand location and the availability of inventory, thus attempting to minimize transportation time while considering inventory availability. This basic approach does not account for the complexity of demand fluctuation or the optimization of the logistics network as a whole. The deterministic optimizer agent, on the other hand, operates based on predefined demand forecasts and attempts to optimize the logistics network under the assumption that future demands will mirror these forecasts. This agent utilizes a standard optimization model to assign commodities to routes, aiming at minimizing the overall transportation and holding cost under static demand conditions. This method, while potentially offering more optimized solutions than the heuristic approach under certain conditions, lacks the flexibility to adapt to real-time changes in demand or supply, making it less effective in the volatile environment of humanitarian logistics.

Unlike the heuristic and deterministic agents, the DRL agent continually learns from the environment through interaction, dynamically adjusting its decisions based on the current state of the logistics network. The integration of GNN allows the DRL agent to better understand and process the complex relationships and dependencies within the logistics network, such as the interconnectedness of routes and the influence of routing decisions on future states of the network. This enables the DRL agent to anticipate changes in demand, adapt to disruptions in supply or transportation, and optimize consignment rerouting in real-time, offering a more robust solution for managing the uncertainties and complexities.

We implemented the simulation environment using the Gymnasium interface provided by the Farma Foundation to facilitate the training and assessment of the DQN agent, as detailed in our methodological approach. This environment enables the systematic simulation of various logistical scenarios, accurately mirroring the complexities of humanitarian aid distribution with a depth and fidelity conducive to rigorous DQN training. Training involved repeated interaction cycles between the DQN agent and the simulated environment, with each cycle or episode representing a sequence of decisions made by the agent that affect the state of the logistics network over a simulated time period. The agent's replay buffer stores experiences from the last 90 steps, initially populated by random samples. A separate neural network copy, which is updated every 100 steps, is used to calculate target values. Furthermore, the agent employs noisy linear layers in all fully connected layers to enhance learning stability and encourage exploration. During training, we utilized the Mean Squared Error (MSE) loss function, an ϵ-greedy exploration strategy with a decaying ϵ parameter, and an Adam optimizer. The experiments were conducted on a computing setup consisting of Intel(R) Core(TM) i7 processors with 16GB RAM and an NVIDIA GeForce RTX 3050 Ti GPU, using Python's "time.process_time()" function to measure the execution time.

4.2 Experimental Results and Analysis

The experimental results reveal distinct differences in performance among the three agents. The DRL agent with GNN architecture consistently outperformed both the greedy heuristic and deterministic optimizer agents across several key metrics. Specifically, the DRL agent demonstrated reduction in the average cost per consignment, with a mean of 247 and a standard deviation of 15.27. This demonstrates the agent's proficiency in cost-saving strategies across various simulated scenarios. In terms of on-time deliveries, the DRL agent achieved a high mean percentage of 94.7% with a standard deviation of 5.47, indicating a consistent and reliable performance in meeting delivery schedules. This is essential in humanitarian contexts where timing is often critical. The frequency of inter-facility transfers was significantly lower for the DRL agent, with a mean of 7 and a standard deviation of 2, suggesting effective anticipation and allocation of supplies that negates the need for frequent transfers. In contrast, traditional methods reflected less optimal outcomes. The deterministic optimization-based policy and the greedy heuristic-based policy reported higher costs, lower on-time delivery percentages, and increased inter-facility transfers. Notably, the DRL agent's standard deviations were lower compared to these traditional methods, suggesting a more predictable and stable performance. We see in Table 1 the overview of the experimental results based on performance metrics for each approach. This economical advantage, coupled with a high consistency in on-time deliveries and a reduced need for inter-facility transfers, underscores the DRL agent's adeptness at optimizing logistics operations dynamically and with high accuracy. The lower standard deviation values across these metrics further articulate the agent's reliability and predictability in performance—a crucial trait for operations in the unpredictable arena of humanitarian efforts.

Table 1: Experimental Results.

	Average Cost per Consignment		**Percentage of On-Time Deliveries**		**Frequency of Inter-Facility Transfers**	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
DRL Agent with GNN	247	15.27	94.7%	5.47	7	2
Deterministic Optimization-Based Policy	315	18.64	91.3%	8.31	12	5
Greedy Heuristic-Based Policy	359	35.88	83.9%	15.76	27	7

Figure 3 illustrates DRL agent's learning curve (on the left figure) and loss reduction over training batches (on the right figure). The upward trend in the progression of average reward per episode over 200 episodes suggests that the DQN agent is learning effectively over time and is consistently improving its policy to maximize the reward. The observed convergence in loss values over 10,000 batches, where a sharp decline is observed initially, which then levels off, indicates that the DRL agent is becoming proficient at predicting the Q-values accurately, which is critical for making informed decisions in the logistics network. These graphical representations corroborate the agent's theoretical advantages, demonstrating not only its learning efficiency but also the stability and reliability of its performance improvements.

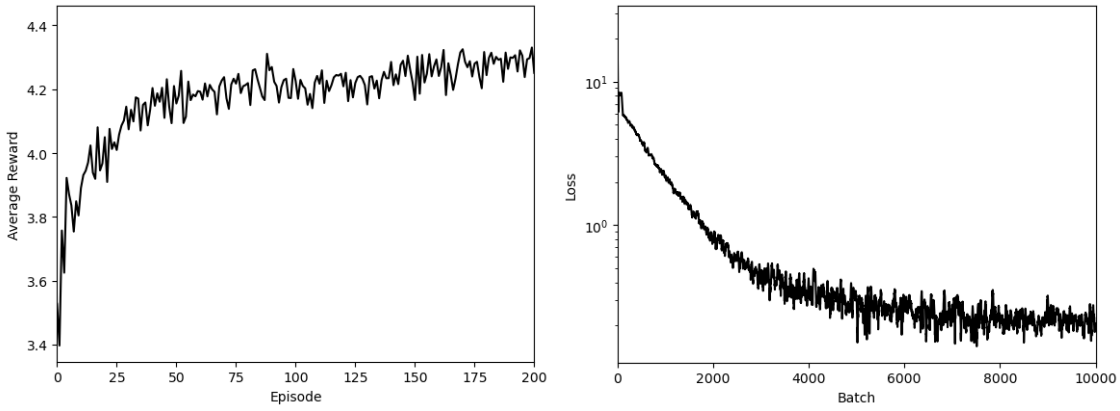

Figure 3: Average Rewards Per Episode and Loss Reduction Over Training Batches for the DQN agent.

4.3 Discussion of Findings

The analysis of experimental results clearly illustrates the innovative potential of the DRL model enhanced by GNN architecture in addressing the optimization challenges of humanitarian logistics. Comparing performance metrics among the three agents - the DRL with GNN, deterministic optimizer, and greedy heuristic - highlights the significant advantages of leveraging advanced machine learning techniques for logistics optimization. Notably, the DRL agent demonstrated superior performance in reducing the average cost per consignment and boosting on-time delivery rates, as well as showcasing operational efficiency through the minimization of inter-facility transfers. These outcomes emphasize the agent's capacity for dynamic adaptation and optimization, essential in the unpredictable and rapidly changing context of humanitarian logistics. The analysis also sheds light on the limitations of traditional optimization methods in the face of the complexities and stochastic nature of humanitarian logistics challenges, pointing to a need for a shift towards more advanced, flexible, and adaptive optimization methods. Furthermore, graphical representations of the DRL agent's learning curve and loss reduction over training batches provide empirical evidence of the model's learning efficiency, stability, and its ability to continuously improve, demonstrating its practical efficacy in optimizing complex logistics networks. The integration of DTs in our approach enhances the realism and fidelity of the simulation environment, offering a robust platform for training and evaluating the DRL agent. By simulating real-world logistics operations and dynamically incorporating changes in the network, the DT enables the DRL agent to train under conditions that closely mimic actual humanitarian logistics scenarios. This integration facilitates the development of highly effective optimization strategies and ensures that the solutions are practical, adaptable, and directly applicable to real-world challenges.

5 CONCLUSION AND FUTURE WORK

This paper demonstrates the efficacy of a DRL approach, improved with the capabilities of DTs, in enhancing the efficiency and adaptability of aid distribution in disaster-affected areas. The results show that our approach, which leverages a DQN algorithm combined with graph-based data processing, surpasses traditional models in optimizing logistical operations. The incorporation of DTs has yielded a robust simulation environment for real-time analysis and scenario planning, fostering proactive and informed decision-making in complex and unpredictable crisis environments. The findings underscore the potential of these integrated technologies to improve resource allocation, ensure timely delivery of aid, and reduce logistical costs. Future research directions include scaling the approach to larger and more varied humanitarian scenarios, integrating Internet of Things (IoT) technology for enhanced data precision, and advancing the predictive analytics capabilities of DTs. Emphasizing human-centered design is essential to ensure these solutions align with the needs of stakeholders in humanitarian logistics. Additionally, incorporating sustainability metrics could align optimization efforts with environmental and social governance goals. Evaluating the effects of policy changes on the performance of humanitarian logistics through simulated models could also provide invaluable insights for policymakers and aid organizations.

REFERENCES

[1] A. Cozzolino and A. Cozzolino, *Humanitarian logistics and supply chain management*. Springer, 2012.

[2] J. M. Day, S. A. Melnyk, P. D. Larson, E. W. Davis, and D. C. Whybark, "Humanitarian and disaster relief supply chains: a matter of life and death," *Journal of Supply Chain Management,* vol. 48, no. 2, pp. 21-36, 2012.

[3] G. Kovács and K. Spens, "Identifying challenges in humanitarian logistics," *International Journal of Physical Distribution & Logistics Management,* vol. 39, no. 6, pp. 506-528, 2009.

[4] M. T. Rahman, T. A. Majchrzak, and T. Comes, "Deep uncertainty in humanitarian logistics operations: decision-making challenges in responding to large-scale natural disasters," *International journal of emergency management,* vol. 15, no. 3, pp. 276-297, 2019.

[5] M. Çelik *et al.*, "Humanitarian logistics," in *New directions in informatics, optimization, logistics, and production*: INFORMS, 2012, pp. 18-49.
[6] L. Özdamar and M. A. Ertem, "Models, solutions and enabling technologies in humanitarian logistics," *European Journal of Operational Research,* vol. 244, no. 1, pp. 55-65, 2015.
[7] T. C. Hu, "Multi-commodity network flows," *Operations research,* vol. 11, no. 3, pp. 344-360, 1963.
[8] A. Hall, S. Hippler, and M. Skutella, "Multicommodity Flows over Time: Efficient Algorithms and Complexity," Berlin, Heidelberg, 2003: Springer Berlin Heidelberg, in Automata, Languages and Programming, pp. 397-409.
[9] D. R. Fulkerson and L. R. Ford, *Flows in networks*. Princeton University Press Princeton, 1962.
[10] A. Haghani and S.-C. Oh, "Formulation and solution of a multi-commodity, multi-modal network flow model for disaster relief operations," *Transportation Research Part A: Policy and Practice,* vol. 30, no. 3, pp. 231-250, 1996, doi: https://doi.org/10.1016/0965-8564(95)00020-8.
[11] S. Lekhavat and S. Chhun, "Multi-commodity supply chain network design problem considering production cost and inventory holding cost," *Journal of Engineering and Innovation,* vol. 15, no. 2, pp. 11-24, 2022.
[12] I. M. Hezam, M. K. Nayeem, and G. M. Lee, "A systematic literature review on mathematical models of humanitarian logistics," *Symmetry,* vol. 13, no. 1, p. 11, 2020.
[13] K. Salimifard and S. Bigharaz, "The multicommodity network flow problem: state of the art classification, applications, and solution methods," *Operational Research,* pp. 1-47, 2022.
[14] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*. MIT press Cambridge, 1998.
[15] P. Almasan, J. Suárez-Varela, K. Rusek, P. Barlet-Ros, and A. Cabellos-Aparicio, "Deep reinforcement learning meets graph neural networks: Exploring a routing optimization use case," *Computer Communications,* vol. 196, pp. 184-194, 2022.
[16] S. Munikoti, D. Agarwal, L. Das, M. Halappanavar, and B. Natarajan, "Challenges and opportunities in deep reinforcement learning with graph neural networks: A comprehensive review of algorithms and applications," *IEEE Trans. on Neural Networks and Learning Systems,* 2023.
[17] M. W. Grieves, "Digital Twins: Past, Present, and Future," in *The Digital Twin*: Springer, 2023, pp. 97-121.
[18] W. Yang, W. Xiang, Y. Yang, and P. Cheng, "Optimizing federated learning with deep reinforcement learning for digital twin empowered industrial IoT," *IEEE Transactions on Industrial Informatics,* vol. 19, no. 2, pp. 1884-1893, 2022.
[19] D. Ivanov, "Intelligent digital twin (iDT) for supply chain stress-testing, resilience, and viability," *International Journal of Production Economics,* p. 108938, 2023.

AUTHOR BIOGRAPHIES

BULENT SOYKAN is a Postdoctoral Researcher at the School of Modeling, Simulation, and Training, University of Central Florida. He received his Ph.D. from Old Dominion University. His research interests include combinatorial optimization, digital twins, reinforcement learning. His email address is bulent.soykan@ucf.edu.

GHAITH RABADI is a Professor and Graduate Programs Director at the School of Modeling, Simulation, and Training, University of Central Florida. He received his Ph.D. in Industrial Engineering from University of Central Florida. His research interests include complex systems modeling, logistics, supply chains, optimization, simulation and AI to find optimal and near-optimal solutions. His email address is ghaith.rabadi@ucf.edu.